ATTORNEY'S DOCKET NO.
TI-32328

APPLICATION FOR UNITED STATES PATENT
DECLARATION AND POWER OF ATTORNEY

As a below named inventor, I declare that my residence, post office address and citizenship are as stated below next to my name; that I verily believe that I am the original, first and sole inventor if only one name is listed below, or an original, first and joint inventor if plural inventors are named below, of the subject matter which is claimed and for which a patent is sought on the invention entitled as set forth below, and the title as set forth below which is described in the attached specification; that I have reviewed and understand the contents of the specification, including the claims, as amended by any amendment specifically referred to in the oath or declaration; that no application for patent or inventor's certificate on this invention has been filed by me or my legal representatives or assigns in any country foreign to the United States of America prior to the filing date of said application; and that I acknowledge my duty to disclose information which is material to the patentability of this application in accordance with Title 37, Code of Federal Regulations, section 1.56;

I further declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true, and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under section 1001 of Title 18 of the United States Code, and that such willful false statements may jeopardize the validity of the application or any patent issuing thereon.

TITLE OF INVENTION:

Method of Speech Recognition with Compensation for Both Channel Distortion and Background Noise

POWER OF ATTORNEY: I HEREBY APPOINT PRACTITIONERS AT CUSTOMER NUMBER **23494** TO PROSECUTE THIS APPLICATION AND TRANSACT ALL BUSINESS IN THE PATENT AND TRADEMARK OFFICE CONNECTED THEREWITH

SEND CORRESPONDENCE TO:

Robert L. Troike
Texas Instruments Incorporated
P.O. Box 655474, MS 3999
Dallas, TX 75265

DIRECT TELEPHONE CALLS TO:

Robert L. Troike
202/639-7710

NAME OF INVENTOR: (1) Yifan Gong	NAME OF INVENTOR: (2) N/A	NAME OF INVENTOR: (3) N/A
RESIDENCE & POST OFFICE ADDRESS: 2504 Trailwest Lane Plano, Texas 75025	RESIDENCE & POST OFFICE ADDRESS:	RESIDENCE & POST OFFICE ADDRESS:
COUNTRY OF CITIZENSHIP: France	COUNTRY OF CITIZENSHIP:	COUNTRY OF CITIZENSHIP:
SIGNATURE OF INVENTOR: x	SIGNATURE OF INVENTOR:	SIGNATURE OF INVENTOR:
DATE: x 22mar, 2001	DATE:	DATE:

Method of Speech Recognition with Compensation for both Channel Distortion and Background Noise

Field of Invention

This invention relates to speech recognition and more particularly to compensation for both background noise and channel distortion.

Background of Invention

A speech recognizer trained with relatively a quiet office environment speech data and operating in a mobile environment may fail due to at least to the tow distortion sources of back ground noise and microphone changes. The background noise may be from a computer fan, car engine, and/or road noise. The microphone changer may be due to the quality of the microphone, hand-held or hands-free and, a position to the mouth. In mobile application of speech recognition, both the microphone conditioner and background noise are subject to change.

Cepstral Mean Normalization (CMN) removes utterance mean and is a simple and effective way of dealing with convolutive distortion such as telephone channel distortion. See "Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification" of B. Atal in Journal of Acoustics Society of America, Vol. 55: 1304–1312, 1974. Spectral Subtraction (SS) reduces background noise in the feature space. See article "Suppression of Acoustic Noise in Speech Using Spectral Subtraction" of S.F. Boll in IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP-27(2): 113-129, April 1979. Parallel Model Combination (PMC) gives an approximation of speech models in noisy condition from noise-free speech models and noise estimates. See "An Improved Approach to the Hidden Markov Model Decomposition of Speech and Noise" of M.J. F. Glaes and S. Young in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Volume 1, pages 233-236, U.S.A., April 1992. The techniques do not require any training data.

Joint compensation of additional noise and convolutive noise can be achieved by the introduction of a channel model and a noise model. A spectral bias for additive noise and a cepstral bias for convolutive noise are introduced in an article by M. Afify, Y. Gong, and J. P. Haton. This article is entitled "A General Joint Additive and Convolutive Bias Compensation Approach Applied to Noisy Lombard Speech Recognition" in IEEE Trans. on Speech and Audio Processing, 6(6): 524-538, November 1998. The five biases can be calculated by application of Expectation Maximization (EM) in both spectral and convolutive domains. A procedure by J.L. Gauvain, et al, is presented to calculate convolutive component, which requires rescanning of training data. See J.L. Gauvain, L. Lamel, M. Adda-Decker, and D. Matrouf entitled "Developments in Continuous Speech Dictation using the ARPA NAB News Task." In Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing, pages 73-76, Detroit, 1996. Solution of the convolutive component by steepest descent method has also been reported. See Y. Minami and S. Furui entitled "A Maximum Likelihood Procedure for a Universal Adaptation Method Based on HMM Composition." See Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing, pages 129-132, Detroit, 1995. A method by Y. Minami and S. Furui needs additional universal speech models, and redestination of channel distortion with the universal models when channel changes. See Y. Minami and S. Furui entitled "Adaptation Method Based on HMM Composition and EM Algorithm" in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, pages 327-330, Atlanta 1996.

The techniques presented by M.F.J. Gales in "PMC for Speech Recognition in Additive and Convolutional Noise," Technical Report TR-154, CUED/F-INFENG, December 1993 needs two passes of the test utterance, e.g., parameter estimation followed by recognition, several transformations between cepstral and spectral domains, and a Gaussian mixture model for clean speech.

Alternatively, the nonlinear changes of both type of distortions can be approximated by linear equations, assuming that the changes are small. Jacobian approach, which models speech model parameter changes as the product of a jacobian matrix and the difference in noisy conditions, and statistical linear approximation are along this direction. See S. Sagayama, Y.

Yamaguchi, and S. Takahashi entitled "Jacobian Adaptation of Noisy Speech Models," in Proceedings of IEEE Automatic Speech Recognition Workshop, pages 396-403, Santa Barbara, CA, USA, December 1997. IEEE Signal Processing Society. Also see "Statistical Linear Approximation for Environment Compensation" of N.S. Kim, IEEE Signal Processing Letters, 5(1): 8-10, January 1998.

Maximum Likelihood Linear Regression (MLLR) transforms HMM parameters to match the distortion factors. See "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density HMMs" by C.J. Leggetter and P.C. Woodland in Computer, Speech and Language, 9(2): 171-185, 1995. This method is effective for both sources but requires training data and introduces the dependence to the speakers.

Summary of Invention

In accordance with one establishment of the present inventor a new method that Handles simultaneously noise and channel distortions to make a speaker independent system robust to a wide variety of noises and channel distortions.

Description of Drawing

Figure 1 illustrates a speech recognizer according to one embodiment of the present invention; and

Figure 2 illustrates the method of the present invention generating

Description of Preferred Embodiments of the Present Inventions

Referring to Fig. 1 there is illustrated a speech recognizer according to the present invention. The speech is applied to recognizer 11. The speech is compared to Hidden Markov Models (HMM) 13 to recognize the text. The models initially provided on those with speech recorded in a quiet environment and the microphone of good quality. We want to develop a speech model set suitable for operating in the simultaneous presence of channel/microphone distortion and background noise. In accordance with the present invention, a speech model set is provided using statistics about the noise and speech. A low computation cost method integrates both PMC and CMN.

Referring to Figure 2, the first Step 1 is to start with HMM models trained on clean speech, with cepstral mean normalization. We modify these models to get models to compensate for channel/microphone distortion (convolutive distortion) and simultaneous background noise (additive distortion). For an HMM model, we have a lot of parameter but only change one subset of the parameters and that is mean vectors $m_{p,j,k}$. The mean vectors $m_{p,j,k}$ of the original model space is modified where p is the index of the Probability Density Function (PDF), j is the state and k is the mixing component.

The second Step 2 is to calculate which is the mean mel-scaled cesptrum coefficients (MFCC) vector over the trained database. Scan all data and calculate the mean to get $\hat{b}$.

The third Step 3 is to add mean $\hat{b}$ to each of this mean vector pool represented by $m_{p,j,k}$ equation (1) to get:

$$\overline{m}_{p,j,k} = m_{p,j,k} + b. \qquad (1)$$

For example, there could be 100 PDFs, 3 states per PDF and 2 vectors per state, or a total of 600 vectors.

The fourth Step 4 is for a given input test utterance, an estimate of the background noise vector $\widetilde{\mathrm{X}}$ is calculated.

Let $\mathrm{u}^l \triangleq [u_1^l, u_2^l, \cdots u_D^l]^T$ and $\mathrm{v}^l \triangleq [v_1^l, v_2^l, \cdots v_D^l]^T$, where l means that the values are represented in log-spectral domain.

We introduce the combination operator $\oplus$ such that:

$$\mathbf{w}^l \triangleq \mathbf{u}^l \oplus \mathbf{v}^l = [w_1^l, w_2^l, \cdots w_D^l]^T \qquad (2)$$

with

$$w_j^l = \log(\exp(u_j^l) + \exp(v_j^l)) \qquad (3)$$

In Step 5, we calculate the mean vectors adapted to the noise $\widetilde{\mathrm{X}}$ using equation 4.

$$\hat{\mathrm{m}}_{p,j,k} = IDFT(DFT(\overline{\mathrm{m}}_{p,j,k}) \oplus DFT(\widetilde{\mathrm{X}})). \qquad (4)$$

where *DFT* and *IDFT* are, respectively, the DFT and inverse DFT operation, $\overline{\mathrm{m}}_{p,j,k}$ is the noise compensated mean vector.

Equation 4 involves several operators. DFT is Discrete Fourier Transform and IDFT is the Inverse Discrete Fourier Transform. The $\oplus$ is an operation with two vectors. $A \oplus B = C$. How $\oplus$ is defined, we look at equations 2 and 3. Equation 2 says the operation + operates on two vectors u and v and the result is a vector of D dimension or $[w_1^l, w_2^l, \cdots w_D^l]^T$ where T is the transposition. We take the two vectors and produce another vector. We need to specify each element in the resultant vector. Equation 3 says that the jth element in that vector (w_j^l) is defined by the exponential of the element of u added to the exponential if the jth element of v and take the log of the combination of the exponential of u added to the exponential of the j the element of v. This completes the definition of Equation 4.

In the following steps, we need to remove the mean vector $\hat{b}$ of the noisy data y over the noisy speech space $\mathcal{N}$ (from the resultant model). One may be able to synthesize enough noisy data from compensated models but this requires a lot of calculation. In accordance with the present invention the vector is calculated using statistics of the noisy models. The whole recognizer will operate with CMN (cepstral mean normalization mode), but the models in Equation 4 are no longer mean normalized. We have dealt with additive noise. The second half of the processing is removing the cepstral mean of our models defined in Equation 4. This is not difficult because we have the models in Equation 4. In Step 6, we need to integrate all the samples generated by Equation 4 to get the mean. Mean is $\hat{b}$. Equation 5 is this integration.

Let , be the variable denoting PDF index, $\mathcal{J}$ be the variable for state index, and $\mathcal{K}$ be the variable for mixing component index.

$$\hat{b} = E\{y\} \tag{5}$$

$$= \int_{\mathcal{N}} y \sum_{p} \sum_{j} \sum_{k} P_{,}(p) P_{\mathcal{J}|,}(j,p) P_{\mathcal{K}|,,\mathcal{J}}(k \mid p,j) p_{\mathbf{Y}|,,,\mathcal{J},\mathcal{K}}(\mathbf{y} \mid p,j,k) d\mathbf{y}$$

Since

$$p(y \mid p,j,k) = \mathbf{N}(\mathbf{y}, IDFT(DFT(\overline{m}_{p,j,k}) \oplus DFT(\widetilde{X})), \_\_{p,j,k}) \tag{6}$$

We have

$$\hat{b} = \sum_{p} \sum_{j} \sum_{k} P_{,}(p) P_{\mathcal{J}|,}(j \mid p) P_{\mathcal{K}|,,\mathcal{J}}(k \mid p,j)\, \hat{m}_{p,j,k} \tag{7}$$

Equation 7 shows that $\hat{b}$ can be worked out analytically, and it is not necessary to do the physically generation and integration. The final result is Equation 7 which is the integration into several sums. Sums over probability density functions and the sum over states and sum over mixing components. Then you have several quantities. The $P_{\mathcal{H}}$ is the probability of having the PDF index. The $P_{\mathcal{J}}$ given $\mathcal{H}$ is the probability of being in the state if given the PDFp. The next is the probability the mixing component p,j given we have the PDF index. The mean vector of the compensated mode. To make this complete we remove this $\hat{b}$ from the compensated model to get the target model. This is Step 7. The target model is:

$$\dot{m}_{p,j,k} = \hat{m}_{p,j,k} - \hat{b} \qquad (8)$$

This is what we want to load into our recognizer. This operation is done for each utterance.

Calculation of $\hat{b}$ thus requires the knowledge of probability of each PDF. There are two issues with P $(, = p)$:

- It needs additional storage space.

- It is dependent of the recognition task e.g. vocabulary, grammar.

Although it is possible to obtain that probability, we want to consider the following simplified cases.

This operation to calculate the $\hat{b}$ can be simplified with three approximations. The first one uses equal probabilities for $P_{\mathcal{H}}(p)$ or constraint C.

1. Use equal probabilities for $P_{\mathcal{H}}(p)$:

$$P_{\mathcal{H}}(p) = C \qquad (9)$$

2. Use equal probabilities for $P_{\mathcal{H}}(p)$, $P_{\mathcal{J}\,|}, P(j\,|p)$ and $P_{\mathcal{K}|_{,,}\mathcal{J}}(k\,|\,h,j)$.

$$\begin{aligned} P_{\mathcal{H}}(p) &= \mathrm{C} \\ P_{\mathcal{J}|H}(j\,|p) &= \mathrm{D} \\ P_{\mathcal{K}|H,,\mathcal{J}}(k\,|p,j) &= \mathrm{E} \end{aligned} \tag{10}$$

3. In fact, the case described in Eq-10 consists in averaging the compensated mean vectors $\overline{\mathrm{m}}_{p,j,k}$. Referring to Eq-4 and Eq-1, it can be expected that the averaging reduces the speech part $\mathrm{m}_{p,j,k}$ just as CMN does. Therefore, Eq-7 could be further simplified into:

$$\hat{\mathrm{b}} = IDFT\,(DFT(\mathbf{b}) \oplus DFT\,(\widetilde{\mathrm{X}})). \tag{11}$$

The model $\mathrm{m}^{T}_{p,j,k,t}$ is then used with CMN on noisy speech. Unfortunately, $\hat{\mathrm{b}}$ is a function of both channel and background noise in all above cases. In other words, in presence of noise, there is no guarantee that the channel will be removed by such a vector, as is for CMN.

A subset of WAVES database containing hands-free recordings was used, which consists of three recording sessions: parked-trn (car parked, engine off), parked (car parked, engine off), and city-driving (car driven on a stop and go basis).

In each session, 20 speakers (10 male) read 40 sentences each, giving 800 utterances. Each sentence is either 10, 7 or 4 digit sequence, with equal probabilities. The database is sampled at 8kHz, with MFCC analysis frame rate of 20ms. Feature vector consists of 10 statis and 10 dynamic coefficients.

HMMs used in all experiments are trained in TIDIGITS clean speech data. Utterance-based cepstral mean normalization is used. The HMMs contain 1957 mean vectors, and 270 diagonal variances. Evaluated on TIGIDIT test set, the recognizer gives 0.36% word error rate.

To improve the performance in noisy environment, the variances of the Gaussian PDFs can be MAP adapted to some slightly noisy data, e.g. WAVES parked_eval data. Such adaptation will not affect recognition of clean speech, but will reduce variance mismatch between HMMs the noisy speech.The new algorithm is referred to as JAC (joint compensation of additive noise and convolutive distortion).

	PARKED	DRIVING
BASELINE	1.38	30.3
CMN	0.59	18.8
PMC	1.74	6.29
JAC FULL PROB (Eq-7)	0.84	1.93
JAC CONST PDF PROB (Eq-9)	0.84	2.00
JAC AVERAGE (Eq-10)	1.16	2.18
JAC SIMPLIFIED (Eq-11)	0.48	1.91

Table 1: word error rate (WER) (%) as function of driving conditions and compensations methods

Table–1shows that:

- Compared to noise-free recognition (WER) (0.36%), without any compensation (BASELENE) the recognition performance degrades severely.
- CMN effectively reduces the WER for parked data, but is not effective for driving conditions where additive noise becomes dominant.
- PMC substantially reduces the WER for driving conditions, but gives poor results for parked data where microphone mismatch is dominant.
- All JAC cases give lower WER than non-JAC methods.

- Simplifying Eq-7 to Eq-9 then to Eq-10 results in progressive increase in WER, although the degradation is not severe. Especially, information in PDF probability is not critical to the performance.

- Simplified JAC gives lowest WER in all tests. For this hands-free speech recognition, the new met reduces word error rate by 61% for parked condition and 94% for city driving condition.